1. **Review Existing Unstructured Data and Diagram a New Structured Relational Data Model**

Given 3 Json files Users, Brands, and Receipts.

First, I read these 3 files using python and to be specific using pandas dataframe.

Then I found the following, for users data frame.  
A screenshot of a computer screen

Description automatically generated

Figure users dataframe sample before cleaning

For “\_id” column I found that every row contains an object called ‘$id’ that holds the data of the id same as last Login column.

For brands dataframe

A screenshot of a computer

Description automatically generated

Figure Brands dataframe sample before cleaning

The same here for “\_id” column, but the “cpg” column had 2 objects ref object and id object so I dealt with it as multivalued attribute and separated them into 2 columns.

For Receipts dataframe

A screenshot of a computer

Description automatically generated

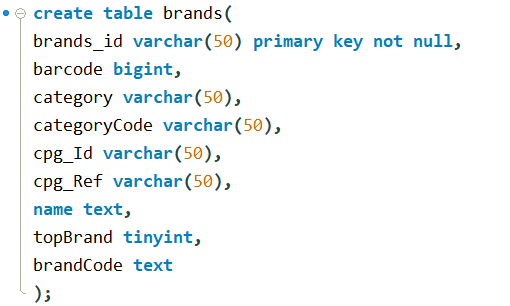
Figure Receipts dataframe sample before cleaning

I found here RewardsReceiptItemList as list of objects and it’s not good to deal with, so we need to recreate new table for this list and link it with the receipts and brands tables.

Next, I created ER diagram using MYSQL workbench, but first created 4 tables with new design and linked them using primary key and foreign key concepts.

Users table  
A screenshot of a computer code

Description automatically generated

Brands table  
****

Receipt Table

A screenshot of a computer code

Description automatically generated

Receipts List Table

A screen shot of a computer code

Description automatically generated

**ERD design**   
A screenshot of a computer

Description automatically generated

In this schema, tables are linked using foreign key references, forming relationships between entities. For example, the receipts\_list table references both the brands table and the receipts table through foreign keys.

1. **Write a query that directly answers a predetermined question from a business stakeholder.**

-- What are the top 5 brands by receipts scanned for most recent month?

-- How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?

-- Query for the Recent Month:

SELECT b.name AS BrandName, COUNT(r.receipts\_id) AS ReceiptsScanned

FROM brands b

JOIN receipts\_list rl ON b.brands\_id = rl.brands\_id

JOIN receipts r ON rl.receipts\_id = r.receipts\_id

WHERE YEAR(r.dateScanned) = YEAR(CURRENT\_DATE)

AND MONTH(r.dateScanned) = MONTH(CURRENT\_DATE)

GROUP BY b.name

ORDER BY ReceiptsScanned DESC

LIMIT 5;

-- Query for the Previous Month:

SELECT b.name AS BrandName, COUNT(r.receipts\_id) AS ReceiptsScanned

FROM brands b

JOIN receipts\_list rl ON b.brands\_id = rl.brands\_id

JOIN receipts r ON rl.receipts\_id = r.receipts\_id

WHERE YEAR(r.dateScanned) = YEAR(CURRENT\_DATE - INTERVAL 1 MONTH)

AND MONTH(r.dateScanned) = MONTH(CURRENT\_DATE - INTERVAL 1 MONTH)

GROUP BY b.name

ORDER BY ReceiptsScanned DESC

LIMIT 5;

-- When considering average spend from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

SELECT rewardsReceiptStatus, AVG(totalSpent) AS AvgSpend

FROM receipts

WHERE rewardsReceiptStatus IN ('Accepted', 'Rejected')

GROUP BY rewardsReceiptStatus;

-- When considering total number of items purchased from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

SELECT rewardsReceiptStatus, SUM(purchasedItemCount) AS TotalItemsPurchased

FROM receipts

WHERE rewardsReceiptStatus IN ('Accepted', 'Rejected')

GROUP BY rewardsReceiptStatus;

-- Which brand has the most spend among users who were created within the past 6 months?

SELECT b.name AS BrandName, SUM(rl.finalPrice) AS TotalSpend

FROM brands b

JOIN receipts\_list rl ON b.brands\_id = rl.brands\_id

JOIN receipts r ON rl.receipts\_id = r.receipts\_id

JOIN users u ON r.user\_id = u.user\_id

WHERE u.createdDate >= DATE\_SUB(CURRENT\_DATE, INTERVAL 6 MONTH)

GROUP BY b.name

ORDER BY TotalSpend DESC

LIMIT 1;

-- Which brand has the most transactions among users who were created within the past 6 months?

SELECT b.name AS BrandName, COUNT(DISTINCT r.receipts\_id) AS TransactionCount

FROM brands b

JOIN receipts\_list rl ON b.brands\_id = rl.brands\_id

JOIN receipts r ON rl.receipts\_id = r.receipts\_id

JOIN users u ON r.user\_id = u.user\_id

WHERE u.createdDate >= DATE\_SUB(CURRENT\_DATE, INTERVAL 6 MONTH)

GROUP BY b.name

ORDER BY TransactionCount DESC

LIMIT 1;

1. **Evaluate Data Quality Issues in the Data Provided**

Users Data (users\_df):

* Duplicates: 283 duplicated rows were found and removed.
* pull out the data from the objects $id and $Date
* A computer code with text

  Description automatically generated with medium confidence
* Null Values: Null values were present in the 'lastLogin', 'signUpSource', and 'state' columns. A screenshot of a computer

  Description automatically generated
* Data Types: The 'createdDate' and 'lastLogin' columns were adjusted to datetime data types.
* Data Cleaning: Null values in 'lastLogin' were replaced with the mean, and null values in 'signUpSource' and 'state' were replaced with the mode.

A screenshot of a computer code

Description automatically generated

Sample of the cleaned users data  
A screenshot of a computer screen

Description automatically generated

Brands Data (brands\_df):

* Duplicates: No duplicated rows were found. A close-up of a computer code

  Description automatically generated
* Null Values: Null values were present in the 'category', 'categoryCode', 'topBrand', and 'brandCode' columns.

A screenshot of a computer code

Description automatically generated

* Data Types: The 'topBrand' column was converted to a boolean data type.
* Data Cleaning: Null values in 'topBrand', 'category', 'categoryCode', and 'brandCode' were replaced with the mode.

A screenshot of a computer

Description automatically generated

Sample of the cleaned Brands data  
A screenshot of a computer screen

Description automatically generated

Receipts Data (receipts\_df):

* Duplicates: No duplicated rows were found. A close-up of a computer screen

  Description automatically generated
* Null Values: Null values were present in several columns, including 'bonusPointsEarned', 'bonusPointsEarnedReason', 'finishedDate', 'pointsAwardedDate', 'pointsEarned', 'purchaseDate', 'purchasedItemCount', and 'totalSpent'.

A screenshot of a computer

Description automatically generated

* Data Types: Several date-related columns were adjusted to datetime data types.
* Data Cleaning: Null values were replaced with the mean for numeric columns and mode for categorical columns. A screenshot of a computer code

  Description automatically generated

**4. COMMUNICATE WITH STAKE HOLDERS**

Subject: Update on Data Quality and Optimization Progress

Hi [STAKEHOLDER],

I hope you’re doing well. I want to give you a quick rundown of the progress we've achieved in advancing our data quality and optimization activities. This message is intended to keep you informed, create a thorough awareness of our work, and identify any potential issues we may be facing.

**Questions About Data Quality**:

While reviewing the data, I encountered some critical questions:

- Is all the data complete, or are there any gaps?

- Can we confidently rely on the accuracy of the data?

- Are we aware of the data sources and any associated concerns?

- Does the data still align with our objectives, or have there been changes?

**Data Quality Issues Detected:**

During the analysis, I observed certain issues that are as follows:

- There were Inconsistencies in numerical values.

- Many redundant records were present that could compromise our insights if not resolved.

- Presence of missing values was high. Though we have replaced it with the mean, it could impact the accuracy of the findings.

**Proposed Solutions for Data Quality Issues:**

To address these concerns, I recommend the following actions:

- Conduct a thorough review of data sources to identify and rectify inconsistencies or inaccuracies.

- Evaluate unusual numerical values and determine if the adjustments are necessary.

- Perform data cleaning to remove the duplicate entries and ensure data integrity.

- Document the data sources and any modifications made.

**Enhancing Data for Optimal Use:**

To further enhance the usability of our data, I'll need the following information:

- Clear understanding of our specific objectives to work on the data accordingly.

- Insights into how the data will be used to optimize its structure and content.

**Anticipating Performance and Scaling Challenges:**

we anticipate the following challenges:

- Increasing data volume could impact on our current infrastructure.

- quick data retrieval may be necessary for timely decision-making.

- Consideration of scalable systems might be required to accommodate growth.

**Mitigating Performance Concerns:**

To address these potential challenges, I'll be focusing on:

- Finding a solution to handle the increase in data volumes.

- Exploring methods to enhance data retrieval for quicker insights.

- Focusing on the need for scalable solutions to support and accommodate future demands.

I'm dedicated to resolving these issues and collaborating with the team to ensure high-quality data. Your insights and input are invaluable in guiding our approach.

Looking forward to discussing this further.

Best regards,

Analytics Team